Efficient Estimation of Conditional Variance Functions in Stochastic Regression

Jianqing Fan*
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599, USA

Qiwei Yao†
Institute of Mathematics and Statistics
University of Kent at Canterbury
Canterbury, Kent CT2 7NF, UK

Abstract

Conditional heteroscedasticity has been often used in modelling and understanding the variability of statistical data. Under a general setup which includes the nonlinear time series model as a special case, we propose an efficient and adaptive method for estimating the conditional variance, therefore also for estimating the volatility function. The method is data-analytic and model free. The basic idea is to apply a local linear regression to the squared residuals. We demonstrate that without knowing the regression function, we can estimate the conditional variance asymptotically as well as if the regression were given. This asymptotic result, established under the assumption that the observations are made from a strictly stationary and absolutely regular process, is also verified via extensive simulations. Further, the asymptotic result paves the way for adapting an automatic bandwidth selection scheme. Various applications with real data sets illustrate the usefulness of our proposed techniques.

Some key words: Absolutely regular; ARCH; Conditional variance; Efficient estimator; Heteroscedasticity; Local linear regression; Nonlinear time series; Volatility.

*Supported partially by NSF Grant DMS-9504414 and NSA Grant 96-1-0015

†Supported partially by EPSRC Grant L16385.

1 Introduction

Many scientific studies depend on understanding the local variability of the data, which is often featured as the conditional variance or the volatility function in a statistical model. It is of common interest to estimate conditional variance functions in a variety of statistical applications such as measuring the volatility or risk in finance (Gallant and Tauchen 1995, Härdle and Tsybakov 1996), monitoring the reliability in nonlinear prediction (Yao and Tong 1994), identifying homoscedastic transforms in regression (Carroll and Ruppert 1988), choosing optimal design and understanding residual pattern (Müller and Stadtmüller 1987, Gasser *et al.* 1986), monitoring the signal-to-noise ratios in quality control of experimental design (Box 1988) and so on. The problem can be mathematically formulated as follows.

Let $\{(Y_i, X_i)\}$ be a two-dimensional strictly stationary process having the same marginal distribution as (Y, X). Let $m(x) = E(Y|X = x)$ and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$ be respectively the regression function and the conditional variance, and $\sigma^2(.) \neq 0$. We write a regression model of Y_i on X_i as

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i. \tag{1.1}$$

Then $E(\epsilon_i|X_i) = 0$ and $\mathrm{Var}(\epsilon_i|X_i) = 1$, although the conditional distribution of ϵ_i given $X_i = x$ may still depend on x. For $X_i = Y_{i-1}$, (1.1) is an autoregressive conditional heteroscedastic (ARCH) time series model, and $\sigma(.)$ is called the volatility function (Engle 1982). The aim of this paper is to derive an efficient fully-adaptive procedure for estimating $\sigma^2(.)$. This approach also allows us to validate some postulated structural models in financial time series analysis (Anderson and Lund 1996b, Gallant and Tauchen 1995). We proceed with the general model (1.1) in this paper, which sometimes also facilitates modelling time series data. (See Examples 1, 2 in §4.1 below.)

There has been substantial literature on the estimation of the nonparametric regression function, yet considerably less attention has been paid to the estimation of conditional variance. Due to the simple decomposition $\sigma^2(x) = E(Y^2|X = x) - \{E(Y|X = x)\}^2$, we tend to use the following obvious and direct estimator in practice:

$$\hat{\sigma}_d^2(x) = \hat{\nu}(x) - \{\hat{m}(x)\}^2, \tag{1.2}$$

where $\hat{m}(.)$ and $\hat{\nu}(x)$ are respectively a regression estimator for $m(.)$ and $\nu(x) \equiv E\{Y^2|X = x\}$. (See, *e.g.* Yao and Tong 1994, Härdle and Tsybakov 1996.) In fact, there are some obvious drawbacks in using such an estimator. For example, $\hat{\sigma}_d^2(.)$ is not always non-negative, especially if different smoothing parameters are used in estimating $m(.)$ and $\nu(.)$. Furthermore, such a direct method can create a very large bias (see §3.1 below). Härdle and Tsybakov (1996) recognized these problems and used a common bandwidth and a common kernel to reduce the bias. While their idea is useful, the approach is still not fully adaptive to the unknown regression function $m(.)$. An alternative regression-adaptive approach is to apply the difference-based estimator (Rice 1984, Gasser *et al.* 1986, Müller and Stadtmüller 1987, also §3.2 below), which uses a high-pass filter to remove the regression function from the data sequence $\{Y_i\}$. Hall *et al.* (1990) demonstrated that the resulting estimator was inefficient even in homoscedastic models with optimal filters.

In this paper, we consider a residual-based estimator of the conditional variance. While the idea is not new, see *e.g.* Hall and Carroll (1989) and Neumann (1994), its implications and implementations are novel. In particular, we show that our estimator is fully regression-adaptive in the sense that without knowning $m(.)$, we can estimate the conditional variance function $\sigma^2(.)$ asymptotically as well as if $m(.)$ were known. This phenomenon is observed independently by Ruppert *et al.* (1996) in an i.i.d. setting. Ruppert *et al.* (1996) compute the asymptotic expressions for the conditional mean and variance in the i.i.d setting, while we establish the asymptotic normality in a more general setup including both i.i.d setting and nonlinear time series. These results are complementary each other and they together provide useful insights into the residual-based variance estimation.

The residual-based estimators overcome the bias problem of the method of Härdle and Tsybakov (1996) and reduce the variance of the difference-based estimator. In fact, the residual-based estimators can be considered as generalized difference-based estimators. (See §3.2 below). An interesting consequence of this study is that we do not have to undersmooth the regression function $m(.)$ in order to obtain a regression-adaptive estimator for the conditional variance $\sigma^2(.)$. In practice, this implies that we can use a data-driven bandwidth selector in estimating $m(.)$, then apply the same bandwidth selector with the the squared residuals to estimate $\sigma^2(.)$. Therefore, a fully data-driven procedure can easily be constructed based on a wealth of existing procedures for the local polynomial regression such as the cross-validation procedure, the pre-asymptotic substitution method (Fan and Gijbels, 1995) and the plug-in approach (Ruppert, Sheather and Wand, 1995). This is in marked contrast with the previous methods, where new bandwidth (or filter length) selection problems are encountered. Neumann (1994) reported some interesting results on a residual-based estimator based on the Gasser-Müller kernel regression.

The paper is organized as follows. In §2, we propose and study the residual-based estimator of the conditional variance based on local linear regression. In §3, we compare the performance of our estimator with various procedures in the literature and discuss their mutual relationship. In §4, we present numerical applications with three real data sets and two simulated models. All the technical proofs are given in the Appendix.

2 Main results

2.1 Estimator

A regression-adaptive estimator for the conditional variance function $\sigma^2(\cdot)$ is one that works asymptotically as well as if the regression function $m(\cdot)$ were given. If the regression function $m(\cdot)$ is given, then from the relation

$$E(r|X=x)=\sigma^2, \quad \text{where} \quad r=\{Y-m(X)\}^2,$$

we can regard the problem of estimating $\sigma^2(\cdot)$ as a nonparametric regression problem. Given the observations $\{(Y_i,X_i), i=1,\ldots,n\}$ from model (1.1), the local linear estimator of $\sigma^2(\cdot)$ is $\hat{\sigma}_b^2(x)=\hat{\alpha}$, where

$$(\hat{\alpha},\hat{\beta})=\arg\min_{\alpha,\beta}\sum_{i=1}^{n}\{r_i-\alpha-\beta(X_i-x)\}^2 W\left(\frac{X_i-x}{h_1}\right), \tag{2.1}$$

the subscript b stands for 'benchmark', and $W(.)$ is a density function on R and $h_1>0$ is a bandwidth. See Fan and Gijbels (1996). The local linear estimators have several nice properties. They possess high statistical efficiency in an asymptotic minimax sense and are design-adaptive (Fan, 1993). Further, unlike many other nonparametric methods, they automatically correct edge effects (Fan and Gijbels, 1992; Ruppert and Wand 1994; Hastie and Loader 1995). Therefore, $\hat{\sigma}_b^2(.)$ provides a benchmark to our problem.

In practice, $m(.)$ is typically unknown. A natural approach is to substitute $m(\cdot)$ by a nonparametric regression estimator. We choose the local linear estimator because of its optimal properties mentioned above. Let $\hat{m}(x)=\hat{a}$ be the local linear estimator that solves the following weighted least-squares problem:

$$(\hat{a},\hat{b})=\arg\min_{a,b}\sum_{i=1}^{n}\{Y_i-a-b(X_i-x)\}^2 K\left(\frac{X_i-x}{h_2}\right), \tag{2.2}$$

where $K(.)$ is a density function on R and $h_2>0$ is a bandwidth. Denote the squared residuals by $\hat{r}_i=\{Y_i-\hat{m}(X_i)\}^2$. This leads to the residual-based estimator $\hat{\sigma}^2(x)=\hat{\alpha}$ with kernel W and bandwidth h_1, where

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \{\hat{r}_i - \alpha - \beta(X_i - x)\}^2 W\left(\frac{X_i - x}{h_1}\right). \tag{2.3}$$

Although the above idea appears somewhat ad hoc, it has interesting implications. Specifically, while the bias for $\hat{m}$ itself is of order $O(h_2^2)$, its contribution to $\hat{\sigma}^2(.)$ is only of $o(h_2^2)$. This can intuitively explained as follows: Observe that

$$\hat{r}_i - r_i = 2\{m(X_i) - \hat{m}(X_i)\}\varepsilon_i + \{m(X_i) - \hat{m}(X_i)\}^2,$$

where $\varepsilon_i = Y_i - m(X_i)$. It is intuitively clear tht the biases of the residuals are of order $O\{h_2^4 + (nh_2)^{-1}\}$ and this is the effect of the estimated regression function on the estimated variance. See Theorem 1 and Remark 2. The result also paves the way for adapting a fully data-driven bandwidth procedure in our estimation.

2.2 Asymptotic normality

To discuss the asymptotic properties, we need the following regularity conditions. Denote by $p(.)$ the marginal density function of X. We use c to denote a generic constant which may be different at different places.

(C1) For a given point x, the functions $E\{Y^k|X = z\}$ and $p(z)$ are continuous at the point x for $k = 3, 4$, and $\ddot{m}(z) \equiv \frac{d^2}{dz^2}m(z)$ and $\ddot{\sigma}^2(z) \equiv \frac{d^2}{dz^2}\sigma^2(z)$ are uniformly continuous on an open set containing the point x. Further, assume $p(x) > 0$.

(C2) $E\{Y^{4(1+\delta)}\} < \infty$, where $\delta \in [0, 1)$ is a constant.

(C3) The kernel functions W and K are symmetric density functions each with a bounded support in R. Further, $|W(x_1) - W(x_2)| \leq c|x_1 - x_2|$ for all x_1 and x_2 in the support of the function W, and $|K(x_1) - K(x_2)| \leq c|x_1 - x_2|$ for all x_1 and x_2 in the support of K.

(C4) The strictly stationary process $\{(X_i, Y_i)\}$ is absolutely regular, i.e.

$$\beta(j) \equiv \sup_{i\geq 1} E\left\{\sup_{A\in\mathcal{F}_{i+j}^{\infty}} |P(A|\mathcal{F}_1^i) - P(A)|\right\} \to 0, \quad \text{as } j \to \infty,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $\{(X_k, Y_k) : k = i, \dots, j\}$, $(j \geq i)$. Further, for the same δ as in (C2), $\sum_{j=1}^{\infty} j^2 \beta^{\frac{\delta}{1+\delta}}(j) < \infty$. (We use the convention $0^0 = 0$.)

(C5) As $n \to \infty$, $h_i \to 0$, and $\liminf_{n\to\infty} nh_i^4 > 0$, for $i = 1, 2$.

We impose the boundedness on the supports of $K(.)$ and $W(.)$ for brevity of proofs; it may be removed at the cost of lengthier proofs. In particular, the Gaussian kernel is allowed. The assumption of the convergence rate of $\beta(j)$ is also for technical convenience. The assumption on the convergence rates of h_1 and h_2 is not the weakest possible.

Remark 1. When $\{(X_t, Y_t)\}$ are independent, (C4) holds with $\delta = 0$ and condition (C2) reduces to $E(Y^4) < \infty$. On the other hand, if (C4) holds with $\delta = 0$, there are at most finitely many non-zero $\beta(j)'$s. This means that there exists an integer $0 < j_0 < \infty$ for which (X_i, Y_i) is independent of $\{(X_j, Y_j),\ j \geq i + j_0\}$, for all $i \geq 1$.

Theorem 1. Suppose that conditions (C1) — (C5) hold. Then,

$$\sqrt{nh_1}\{\hat{\sigma}^2(x) - \sigma^2(x) - \theta_n\} \xrightarrow{d} N\left(0,\ p^{-1}(x)\sigma^4(x)\lambda^2(x)\int W^2(t)dt\right),$$

where $\lambda^2(x) = E\{(\epsilon^2 - 1)^2 | X = x\}$, $\epsilon = \{Y - m(X)\}/\sigma(X)$, $\sigma_W^2 = \int t^2 W(t)dt$, and

$$\theta_n = \frac{h_1^2}{2}\sigma_W^2 \ddot{\sigma}^2(x) + o(h_1^2 + h_2^2). \qquad (2.4)$$

Remark 2. In the bias of $\hat{\sigma}^2(x)$, the contribution from the error in the estimator $\hat{m}(x)$ enters as a higher order infinitesimal than h_2^2, namely the order of the bias of $\hat{m}(x)$ itself. This permits us to use the optimal bandwidth to smooth $\hat{m}$ — no undersmooth of $\hat{m}$ is needed. Our proof shows further that the bias in (2.4) is of form

$$\theta_n = \frac{h_1^2}{2}\sigma_W^2 \ddot{\sigma}^2(x) + o(n^{-3/5}),$$

if the bandwidths with optimal rates (*i.e.* $h_1 = O(n^{-1/5})$ and $h_2 = O(n^{-1/5})$) are used.

2.3 Efficiency

It follows from the local linear regression theory (see e.g. §5.4.4 of Fan and Gijbel 1996), the benchmark estimator $\hat{\sigma}_b^2(.)$ derived from (2.1) is asymptotically normal. The leading terms in asymptotic bias and variance are as follows:

$$\text{bias}\{\hat{\sigma}_b^2(x)\} : \frac{h_1^2}{2}\sigma_W^2 \ddot{\sigma}^2(x), \quad \text{Var}\{\hat{\sigma}_b^2(x)\} : \frac{1}{nh_1}\frac{\sigma^4(x)\lambda^2(x)\int W^2(t)dt}{p(x)}. \qquad (2.5)$$

Theorem 1 shows that the asymptotic variance of $\hat{\sigma}^2(x)$ admits the same leading term as that of $\hat{\sigma}_b^2(x)$, while the asymptotic biases of the two estimators are also the same provided h_2 converges to 0 not slower than h_1. This is a very minor requirement. It is well known that the best h_2 for estimating $m(.)$ should be of the order $n^{-1/5}$. Substituting such an h_2 in (2.4), the optimal h_1 which minimizes the asymptotical mean squared error is also of the order $n^{-1/5}$. Therefore, the estimator $\hat{\sigma}^2(.)$ is efficient and adaptive to the unknown regression function $m(\cdot)$.

2.4 Bandwidth selection

Bandwidth parameter is important to virtually any nonparametric estimators. The results given in §2.2 permit us to take advantage of existing bandwidth selection methods for the local linear fit. Let $\hat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ be a data-driven bandwidth selection rule for the local linear regression based on the data $(X_1, Y_1), \cdots, (X_n, Y_n)$. This can be derived in one case by *e.g.* the cross-validation bandwidth rule, and in another case by either the pre-asymptotic substitution method of Fan and Gijbels (1995) or by the plug-in method of Ruppert, Sheather and Wand (1995). The latter two methods have been demonstrated to be less variable and more effective. In all cases, $\hat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ is a consistent estimate of the asymptotic optimal bandwidth, which is of order $O(n^{-1/5})$. Our bandwidth selection rule reads as follows:

1. Use bandwidth $h_2 = \hat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ in local linear regression (2.2) to obtain the estimate $\hat{m}(X_i)$ for $i = 1, \cdots, n$.

2. Compute squared residuals $\hat{r}_i = \{Y_i - \hat{m}(X_i)\}^2$, $i = 1, \cdots, n$.

3. Apply bandwidth $h_1 = \hat{h}(X_1, \cdots, X_n; \hat{r}_1, \cdots, \hat{r}_n)$ in local linear regression (2.3) to obtain $\hat{\sigma}^2(\cdot)$.

In the above algorithm, we keep the bandwidth selection method flexible. In our implementation, we use the pre-asymptotic substitution method by Fan and Gijbels (1995), since it has been demonstrated that the resulting estimator possesses fast relative rate of convergence (Huang, 1995).

3 Other estimators

3.1 Direct estimators

Härdle and Tsybakov (1996) proposed an improved version of the direct estimator $\hat{\sigma}_d^2(.)$, as given in (1.2), with local polynomial regression estimators $\hat{m}(.)$ and $\hat{\nu}(.)$ using the same kernel function and the same bandwidth, where $\hat{\nu}(x)$ is an estimate for $E(Y^2|X=x)$. They also established the asymptotic normality of the estimator. If the local linear estimators are used with kernel $W(.)$ and bandwidth h_1, the leading terms in the asymptotic bias and the asymptotic variance of $\hat{\sigma}_d^2(x)$ are

$$\text{bias}\{\hat{\sigma}_d^2(x)\}: \ \frac{h_1^2}{2}\sigma_W^2[\ddot{\sigma}^2(x) + 2\{\dot{m}(x)\}^2], \quad \text{Var}\{\hat{\sigma}_d^2(x)\}: \ \frac{1}{nh_1}\frac{\sigma^4(x)\lambda^2(x)\int W^2(t)dt}{p(x)}.$$

On comparing this with (2.5), the direct estimator has the same asymptotic variance as the benchmark $\hat{\sigma}_b^2(.)$, but admits one more term in the bias. This extra term $h_1^2\sigma_W^2\{\dot{m}(x)\}^2$ could lead to an adverse effect on the quality of estimation and is not adaptive to unknown regression functions. For example, even when $m(\cdot)$ is a linear function with a large slope, this direct method would have a large bias. Thus, the estimator $\hat{\sigma}^2(.)$ derived from (2.3) appears more appealling.

The existence of one more term in the bias of the direct estimator can be understood through the following heuristic arguments. Note that

$$\hat{\sigma}_d^2(x) - \sigma^2(x) = \{\hat{\nu}(x) - \nu(x)\} - 2m(x)\{\hat{m}(x) - m(x)\} - \{\hat{m}(x) - m(x)\}^2. \tag{3.1}$$

The first term on the RHS has the bias

$$\frac{h_1^2}{2}\sigma_W^2\ddot{\nu}(x) = \frac{h_1^2}{2}\sigma_W^2\ddot{\sigma}^2(x) + h_1^2\sigma_W^2\{\dot{m}(x)\}^2 + h_1^2\sigma_W^2 m(x)\ddot{m}(x), \tag{3.2}$$

in which the last term on the RHS will cancel the bias of the second term on the RHS of (3.1). Note that the bias from the third term on the RHS of (3.1) is of the order h_1^4. Therefore, the term involving $\{\dot{m}(x)\}^2$ stays. This argument also shows that using different kernels or bandwidths in the estimators $\hat{m}(.)$ and $\hat{\nu}(.)$ could further increase the bias of $\hat{\sigma}_d^2(.)$. Note that the inefficiency of the estimator $\hat{\sigma}_d^2(.)$ cannot be rescued by using higher order polynomials in the local fitting or by a fitting with different orders. For example, the last term in the RHS of (3.2) will remain in the bias if the local quadratic smoother was used in estimating $m(.)$ while we retain the local linear estimator for $\nu(.)$.

Why can the residual-based estimator $\hat{\sigma}^2(.)$ give smaller bias? To gain some insight, let us consider the local constant smoother, namely setting β equal 0 in (2.3). Then the resulting estimator is

$$\frac{\sum_{i=1}^n \{Y_i - \hat{m}(X_i)\}^2 W\left(\frac{X_i - x}{h_1}\right)}{\sum_{i=1}^n W\left(\frac{X_i - x}{h_1}\right)}.$$

This estimator will reduce to the direct estimator $\hat{\sigma}_d^2(x)$ if all the $\hat{m}(X_i)'$s in the above expression are replaced by $\hat{m}(x)$. Clearly, $\{Y_i - \hat{m}(x)\}^2$ is more biased for $E\{Y - m(X)\}^2$ than $\{Y_i - \hat{m}(X_i)\}^2$. This explains why the residual-based estimator inherits less bias from $\hat{m}(.)$ than the direct estimator.

3.2 Difference-based estimators

For a fixed design model

$$Y_i = m(x_i) + \sigma(x_i)\epsilon_i,$$

in which $x_1 \leq \ldots \leq x_n$ are fixed, $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = 1$, Müller and Stadtmüller (1987) proposed to estimate $\sigma^2(.)$ through a difference sequence. Their approach can be *briefly* described as follows. Form an initial local variance estimate

$$\tilde{\sigma}^2(x_i) = \left(\sum_{j=-m}^{m} w_j Y_{i+j}\right)^2, \tag{3.3}$$

where $m > 0$ is a prescribed integer, and the difference sequence $\{w_j\}$ satisfies the conditions

$$\sum_{j=-m}^{m} w_j = 0, \quad \sum_{j=-m}^{m} w_j^2 = 1. \tag{3.4}$$

By writing

$$\sigma^2(x_i) = \tilde{\sigma}^2(x_i) + \tilde{\epsilon}_i,$$

a kernel moother is applied to obtain the final estimator for $\sigma^2(.)$ based on the above regression relationship.

Estimators of this type have a long history in the time series context; see, for example, Anderson (1971, p.66). The application in nonparametric homoscedastic regression includes Rice (1984), Gasser *et al.* (1986), and Hall *et al.* (1990). It is shown by Hall *et al.* (1990) that if the optimal difference sequence of $\{w_i\}$ is employed for a Gaussian model, the efficiency of the estimator is $4m/(4m+1)$.

In fact, the residual-based estimator $\hat{\sigma}^2(.)$ can be regarded as a generalized difference-based estimator, and $\hat{r}_i$ serves as a crude estimate of $\sigma^2(X_i)$. To make such a connection, we express the local linear estimator of $m(.)$ as

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i.$$

Then, it can be shown that $\sum_{i=1}^n w_i(x) = 1$. Write

$$\hat{r}_i = \{Y_i - \hat{m}(X_i)\}^2 = \left(\sum_{j=1}^n w_{i,j} Y_i\right)^2,$$

where $w_{i,i} = 1 - w_i(X_i)$ and $w_{i,j} = -w_j(X_i)$ for $i \neq j$. Obviously, $\{w_{i,j}\}$ is a difference sequence satisfying $\sum_{j=1}^n w_{i,j} = 0$. However, such a sequence of $\{w_{i,j}\}$ does not exactly fulfil the second condition in (3.4), but

$$\sum_j w_{i,j}^2 = 1 + O_p\{(nh_2)^{-1}\}.$$

The effective length of the sequence is $2m = 2nh_2$, which tends to infinity. This also explains why the estimator $\hat{\sigma}^2(.)$ is efficient in contrast to such results as Hall *et al.* (1990).

Estimation of variance functions with more general weights was discussed by Müller and Stadtmüller (1993). The rates of convergence for this class of estimators were thoroughly investigated. In particular, Müller and Stadtmüller (1993) find that it requires only very mild smoothness condition on the regression function in order to obtain the iptimal rates for the variance estimation.

3.3 Maximum locally likelihood estimators

If the distribution of ϵ is known, the locally maximum likelihood approach could be more efficient. See §4.9 of Fan and Gijbels (1996) and the references therein. For example, if $\{\epsilon_i\}$ are independent and normal (or even approximately normal), the log likelihood function can be expressed as

$$-\frac{1}{2}\sum_{i=1}^n L(\sigma^2(X_i), Y_i - m(X_i)),$$

where $L(\alpha, y) = \alpha^{-1}y^2 + \log\alpha$. The local maximum likelihood approach with the local constant smoother leads to estimating $\sigma^2(x)$ by $\hat{\alpha}$, where

$$(\hat{a}, \hat{\alpha}) = \arg\min_{a,\alpha} \sum_{i=1}^n L(\alpha, Y_i - a) W\left(\frac{X_i - x}{h_1}\right).$$

It is easy to see that the resulting estimator for $\sigma^2(.)$ is indeed the direct estimator $\hat{\sigma}_d^2(.)$ with both $\hat{m}(.)$ and $\hat{\nu}(.)$ being the local constant estimators.

The approach with the local linear smoother needs to estimate four functions. To make it more tractable, we substitute $m(.)$ directly by its local linear estimator $\hat{m}(.)$, derived from (2.2). Let $\hat{\alpha}$ and $\hat{\beta}$ be the minimizer of the residual-based likelihood function:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^n L\{\alpha + \beta(X_i - x), Y_i - \hat{m}(X_i)\} W\left(\frac{X_i - x}{h_1}\right).$$

Then, the maximum local likelihood estimator is defined by $\hat{\sigma}^2_{ml}(x) = \hat{\alpha}$. The estimator is also residual-based, and therefore we would expect that it is adaptive to all unknown regression functions in a similar way to what $\hat{\sigma}^2(\cdot)$ does. However, $L(\alpha, y)$ is not a convex function of α, it might need some delicate handling to establish the asymptotic normality of $\hat{\sigma}^2_{ml}(x)$. On the other hand, following the approach suggested in §4.9 of Fan and Gijbels (1996), we can derive the approximate bias and variance of the estimator, which entail the same leading terms as those for $\hat{\sigma}^2_b(x)$ given in (2.5). In this sense, the local maximum likelihood estimator is also efficient. Unfortunately, $\hat{\sigma}^2_{ml}(x)$ does not admit an explicit solution. The Newton-Raphson one-step iterative estimator can be constructed, however. We do not go further in this direction, since $\hat{\sigma}^2(.)$ derived from (2.3) is more direct, distribution-free and efficient.

4 Applications and Simulations

In this section, we first apply the adaptive estimator $\hat{\sigma}^2(.)$ derived from (2.3) to three real data sets. The finding from these applications includes the validation of an existing structural model. Then, extensive simulations are carried out to confirm the theoretical claim that the adaptive estimator works almost as well as the ideal estimator $\hat{\sigma}^2_b(.)$ defined in (2.1). We use two simulated models, one with i.i.d observatios and one nonlinear time series, for illustration, for which the exact conditional variances can be evaluated at least numerically.

Throughout this section, the two dashed curves around a solided curve always indicate the one standard deviation above and below the estimated curve. The conditional variance functions are always estimated by the adaptive estimator $\hat{\sigma}^2(.)$ derived from (2.3) unless specified otherwise. We always use the Epanechnikov kernel in our calculation. All bandwidths are selected using the pre-asymtotic substitution method by Fan and Gijbels (1995).

4.1 Applications

Example 1. This example concerns the yields of the three month Treasury Bill from the secondary market rates (on Fridays). The secondary market rates are annualized using a 360-day year of bank interest and quoted on a discount basis. The data consist of 1,735 weekly observations, from January 5, 1962 to March 31, 1995, and are presented in Figure 1(a). The data were previously analyzed by various authors, including Andersen and Lund (1996a, b) and Gallant and Tauchen (1996).

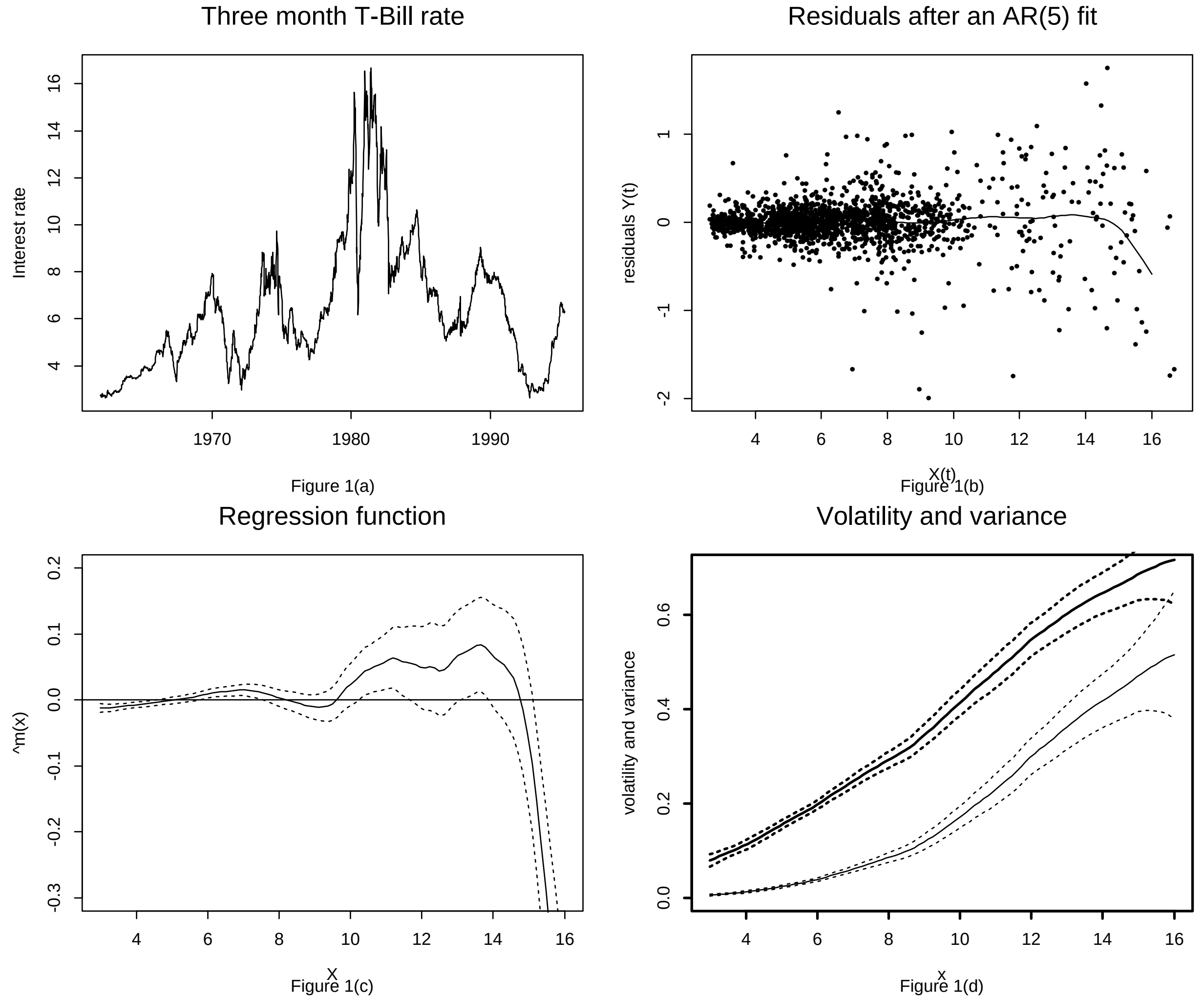

Figure 1(a)

Figure 1(b)

Figure 1(c)

Figure 1(d)

Figure 1: *Three-month Treasure Bill data. (a) Raw data; (b) Residuals after an AR(5) fit is plotted against $X_t \equiv z_{t-1}$; solid curve is the regression curve. (c) The regression curve for the data in (b). (d) The estimated volatility curve (thick curve) and the conditional variance function (thin curve).*

Let z_t denote the time series presented in Figure 1(a). We first fitted an AR model with order selected by the Akaike information criterion. This yields the following AR(5) model:

$$z_t = 1.0733 z_{t-1} - 0.0423 z_{t-2} + 0.0165 z_{t-3} + 0.0228 z_{t-4} - 0.0773 z_{t-5} + Y_t.$$

The selection of AR(5) models coincides with that used by Andersen and Lund (1996b). The ‘residuals’ Y_t are plotted against $X_t \equiv z_{t-1}$ in Figure 1(b). Figure 1(c) depicts the estimated mean regression function $\hat{m}(x) \equiv E\{Y_t | z_{t-1} = x\}$. The nonlinearity with a slightly increasing trend (up to $z_{t-1} = 14$) can be noted. The bandwidth selected by our software is 1.9535. The residual-based estimator for the conditional variance of Y_t given $z_{t-1} = x$ is denoted as $\hat{\sigma}^2(x)$ with the automatically selected bandwidth 3.1461. The estimated volatility function $\hat{\sigma}(x)$ is presented in Figure 1(d). The overall fitted model is

$$z_t = \hat{m}(z_{t-1}) + 1.0733 z_{t-1} - 0.0423 z_{t-2} + 0.0165 z_{t-3} + 0.0228 z_{t-4} - 0.0773 z_{t-5} + \hat{\sigma}(z_{t-1})\epsilon_t,$$

in which $E(\epsilon_t | z_{t-1}) = 0$, and $\mathrm{Var}(\epsilon_t | z_{t-1}) = 1$. Note that the correlation coefficient between the logarithm of z_{t-1} and logarithm of $\hat{\sigma}(z_{t-1})$ is 0.999. This lends a strong support to the structural volatility model

$$\sigma(z_{t-1}) = \alpha z_{t-1}^{\beta},$$

which was considered by Andersen and Lund (1996b). Applying the least-square fit to the log-transformed data, we found that $\alpha = 0.0169$ and $\beta = 1.380$.

Example 2. The data presented in Figure 2(a) are the daily exchange rates (at closing time) between the pounds sterling and US dollars at every business day between 2 March 1980 and 3 April 1993. The length of this time series $\{X_t\}$ is 3,306. To remove the linear trend, we take the difference first. The difference $Y_t = X_{t+1} - X_t$ is plotted against X_t in Figure 2(b). Figure 2(c) shows a nonparametric regression curve of Y_t on X_t, which is denoted by $\hat{m}(X_t)$. Figure 2(d) depicts the volatility curve $\hat{\sigma}(X_t)$. The fitted model is

$$X_{t+1} = X_t + \hat{m}(X_t) + \hat{\sigma}(X_t)\epsilon_t, \quad \text{with} \quad E(\epsilon_t | X_t) = 0, \ \mathrm{Var}(\epsilon_t | X_t) = 1.$$

Figure 2(d) shows clearly that the volatility function for exchange rate is not monotonic. This is in marked contrast with the yields of the Treasure bill data. Ignoring the downward damping on the right edge possibly due to the boundary effect, the volatility curve shows a U-shaped structure, which is also called a 'smiling face'. This implies that the risks of returns are much higher for extreme values taken on the previous day. Härdle and Tsybakov (1996) observed a similar pattern in the exchange rate between German Marks and US dollars.

Example 3. Instead of considering large time series data sets, we now consider a 'cross-sectional' data set. Presented in Figure 3(a) are 133 observations of motorcycle data from Schmidt, Mattern and Schüler (1981). The time (in milliseconds) after a simulated impact on motorcycles was recorded and serves as the covariate X_t. The response variable Y_t is the head acceleration (in gram) of a test object. We fit the data with model (1.1). The estimated regression function $\hat{m}(.)$ is depicted in Figure 3(a). Figure 3(b) describes the residuals and the estimated conditional standard deviation $\hat{\sigma}(.)$. The bandwidths selected for estimating the regression function and

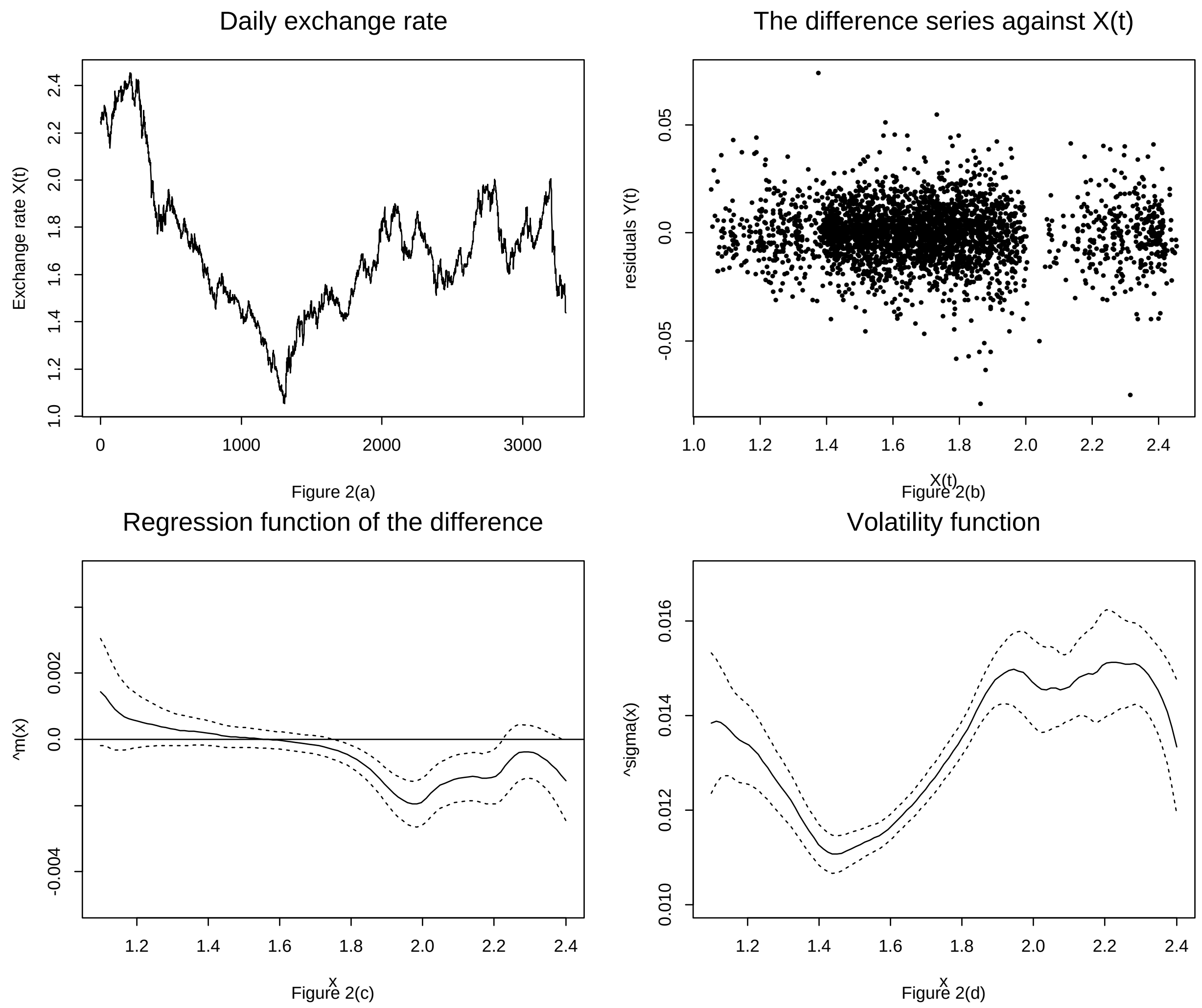

Figure 2(a)

Figure 2(b)

Figure 2(c)

Figure 2(d)

Figure 2: *Exchange rates between the pounds sterling and US dollars. (a) Raw data. (b) The difference series* $\{X_{t+1} - X_t\}$ *is plotted against* X_t. *(c) The regression curve for the data in (b). (d) The estimated volatility curve.*

the conditional variance function are 3.230 and 6.293 respectively. Figure 3(b) shows that $\hat{\sigma}^2(.)$ captures the changes of variability in the data.

4.2 Simulations

Example 4. We simulated 400 random samples of size $n = 200$ from the model

$$Y_i = a\{X_i + 2\exp(-16X_i^2)\} + \sigma(X_i)\epsilon_i, \quad \text{with} \quad \sigma(x) = 0.4\exp(-2x^2) + 0.2,$$

where $\{X_i\}$ and $\{\epsilon_i\}$ are two independent i.i.d. sequences, and $X_i \sim U[-2, 2]$ and $\epsilon_i \sim N(0, 1)$. Four different values of a, namely $a = 0.5, 1, 2, 4$, are used in the simulation. For each simulated sample, the performance of the estimator is evaluated by the Mean Absolute Devation Error

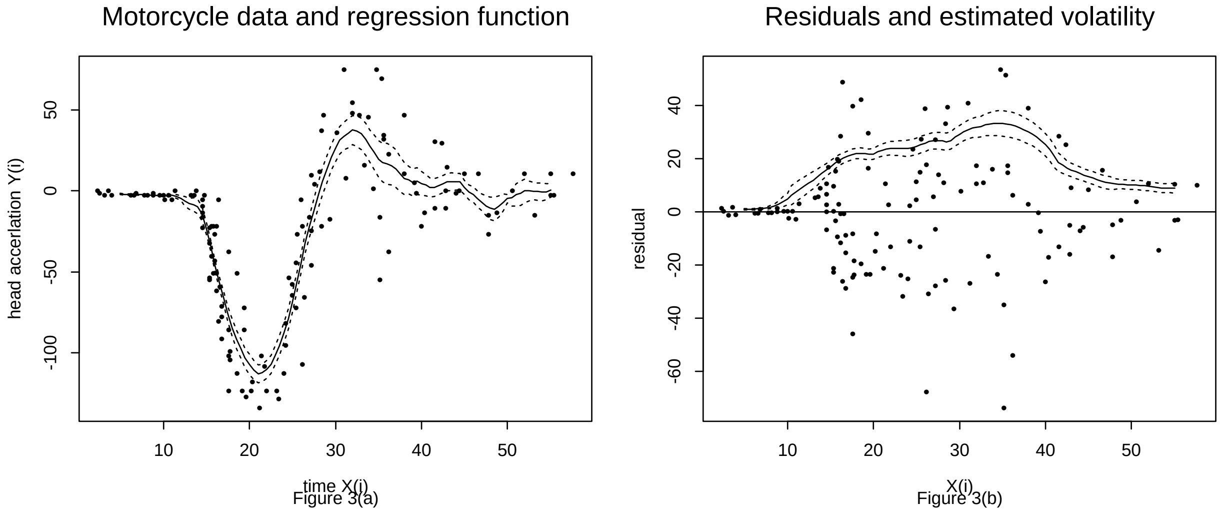

Figure 3: *Motorcycle data (a) Raw data and their estimated regression function. (b) The residuals and the estimated volatility function.*

(MADE):

$$\text{MADE} = n_{\text{grid}}^{-1} \sum_{i=1}^{n_{\text{grid}}} |\hat{\sigma}(x_j) - \sigma(x_j)|,$$

where $\{x_j, j = 1, \cdots, n_{\text{grid}}\}$ are the grid points on $[-1.8, 1.8]$ with $n_{\text{grid}} = 101$. The results are summarized in Figure 4. Figure 4(a) compares the adaptive variance estimator with the ideal variance estimator $\hat{\sigma}_b^2(.)$ which does not vary with different values of a. Presented there are the boxplots of MADEs based on 400 simulations. The first four boxplots are the MADEs of the adaptive estimator $\hat{\sigma}^2(.)$ for $a = 0.5, 1, 2, 4$ in order, and the last one is that of the ideal estimator $\hat{\sigma}_b^2(.)$. As anticipated, the adaptive estimator performs almost as well as the ideal one.

To get further insights, we consider the specific case $a = 1$. The scenario is similar for other cases. Figure 4(b) plots the MADE based on the adaptive estimator versus the MADE based on the ideal estimator, using the same sample data. Clearly, there is about equal chance that one estimator beats the other. The marginal densities of MADE of the adaptive estimator (thick curve) and of the ideal estimator (thin curve) are also depicted in Figure 4(b). This shows again that the performance of the two estimators is comparable. Figure 4(c) presents a typical simulated sample with its corresponding estimated regression function. The typical sample was selected in such a way that the corresponding MADE is equal to its median among the 400 simulations. The sample residuals and the estimated conditional standard deviations are plotted in Figure 4(d). The bandwidths are automatically selected by the procedure outlined in §2.4 and are 0.1867 for the mean regression and 0.4841 for the conditional variance function respectively.

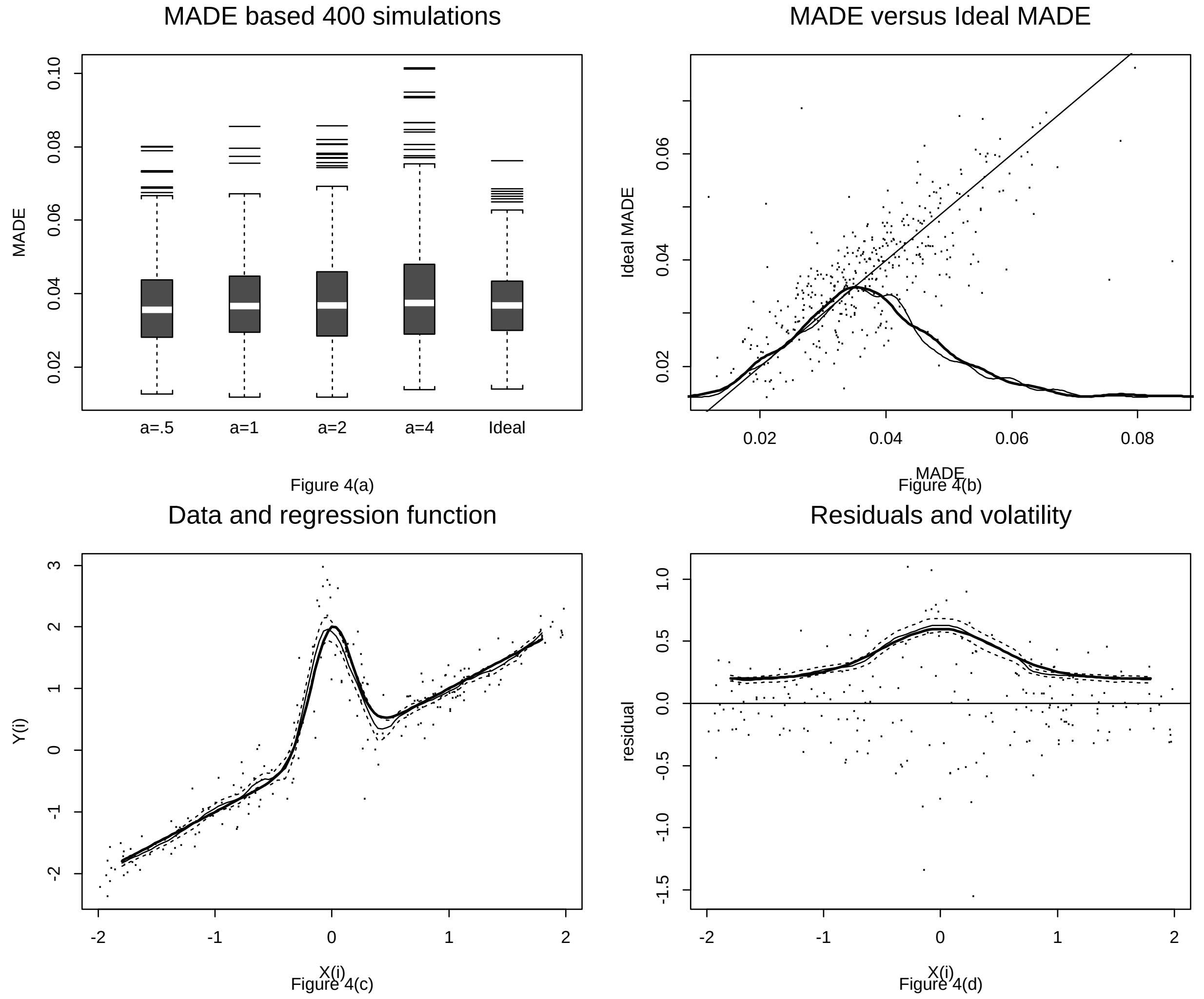

Figure 4: *Simulation results for Example 4. (a) Boxplots of the MADEs for the adaptive estimator with* $a = 0.5, 1, 2, 4$*, and for the ideal estimator (from left to right). (b) The scatter plot of the MADE of* $\hat{\sigma}^2(.)$ *versus the MADE of* $\hat{\sigma}_b^2(.)$*; the straight line marks the position where the two MADEs are equal. Thick curve — the p.d.f. of the MADE of* $\hat{\sigma}^2(.)$*; thin curve — the p.d.f. of the MADE of* $\hat{\sigma}_b^2(.)$*. (c) A representative sample, the corresponding estimated regression curve (thin curve), and the true regression curve (thick curve). (d) The sample residuals from (c), the estimated volatility (thin curve), and the true volatility (thick curve).*

Example 5. Consider the following nonlinear time series:

$$X_{t+1} = 0.235 X_t (16 - X_t) + e_t,$$

where $e_1, e_2, \ldots,$ are independent with the same distribution as $N(0, 0.3^2)$. The skeleton of this model exhibits chaos and has been used by Yao and Tong (1994) to illustrate the influence of the initial values on nonlinear prediction.

For this nonlinear time series, we consider the two-step and three-step forecasting by taking

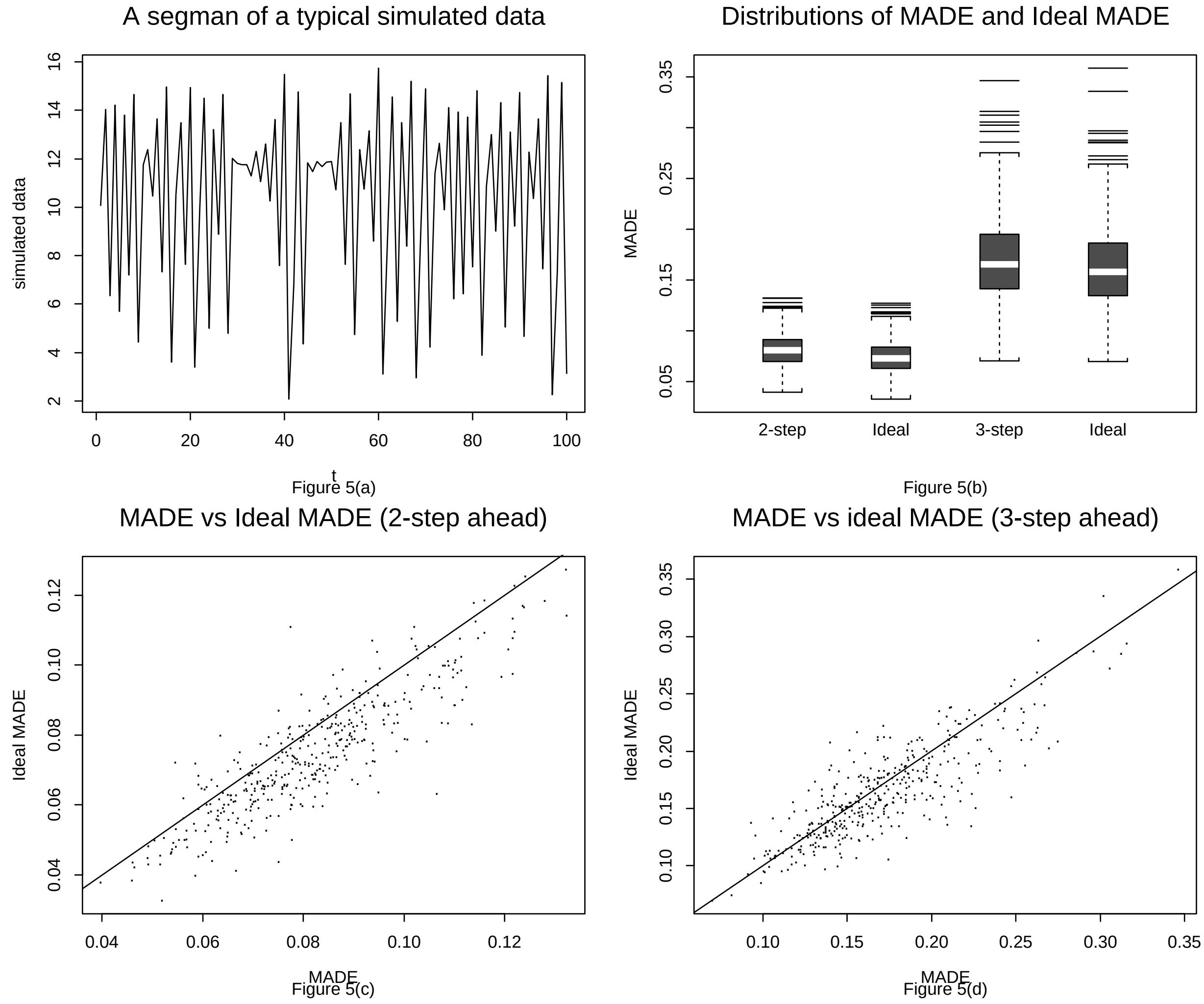
A segman of a typical simulated data

Figure 5(a)

Distributions of MADE and Ideal MADE

Figure 5(b)

MADE vs Ideal MADE (2-step ahead)

Figure 5(c)

MADE vs ideal MADE (3-step ahead)

Figure 5(d)

Figure 5: *Simulation results for Example 5. (a) A fraction of the simulated time series. (b) Boxplots of the MADEs for the adaptive estimator and for the ideal estimator; the left two panels are for 2-step ahead prediction and the right two panels are for the 3-step ahead prediction. (c) The scatter plot of the MADE versus the ideal MADE for 2-step ahead prediction. (d) The scatter plot of the MADE versus the ideal MADE for 3-step prediction. (In both (c) and (d), the straight line marks the position where the two MADEs are equal.)*

respectively $Y_t = X_{t+2}$ and $Y_t = X_{t+3}$. Note that the conditional variance functions concerned are not constant. On the other hand, the conditional variance of the one-step prediction is a constant, and is therefore not presented here.

Figure 5(b) compares the ideal estimator with the adaptive estimator based on 400 simulations with $n = 500$. As we can see, the adaptive estimator works almost as well as the ideal estimator. Figures 5(c) and 5(d) give the scatter plot of MADE for the adaptive estimator and the ideal estimator, using the same sample data. A typical simulated data set and the corresponding

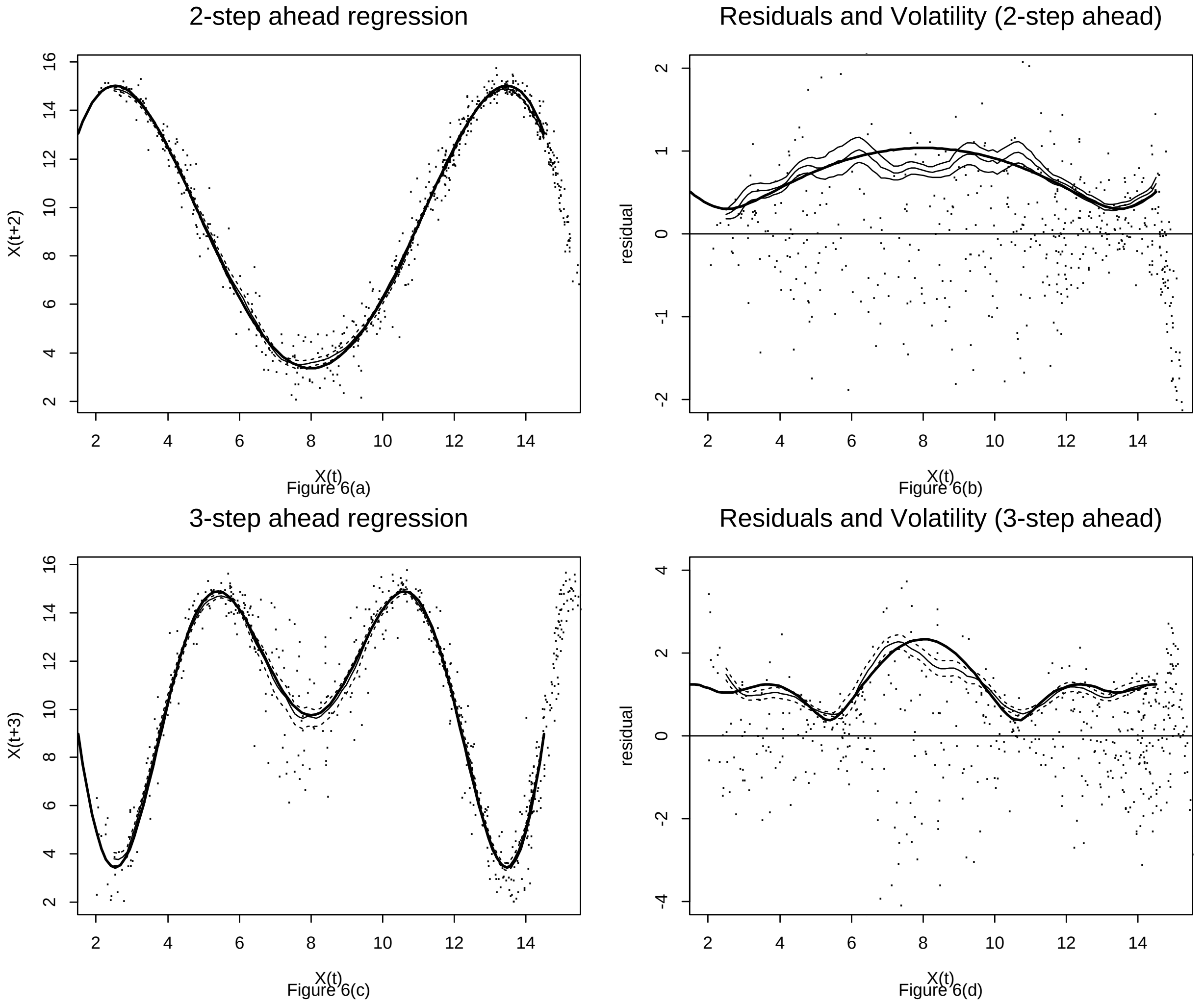

Figure 6(a)

Figure 6(b)

Figure 6(c)

Figure 6(d)

Figure 6: *The results for a representative sample of Example 5. (a) A representative sample and its estimated 2-step ahead regression curve. (b) The sample residuals of (a), the true volatility function (thick curve), and estimated volatility function. (c) A representative sample and its estimated 3-step ahead regression curve. (b) The sample residuals of (c), the true volatility function (thick curve), and estimated volatility function.*

estimated curves are presented in Figure 6. The criterion used to choose a typical sample is again the one for which the MADE is equal to its median among the 400 simulations. Figures 6(a) and 6(c) present the estimated regression functions for 2-step and 3-step ahead prediction respectively, where bandwidths 0.6705 and 0.5577 were selected by our procedure. Their estimated volatility functions are presented in Figures 6(b) and 6(d). The selected bandwidths are 0.6705 and 0.8165 respectively.

Acknowledgement. We would like to thank Professor Ron Gallant for kindly providing us the data analyzed in Example 1.

References

Andersen, T.G. and Lund, J. (1996a). Estimating continuous time stochastic volatility models of the short term interest rate. *Journal of Econometrics*, forthcoming.

Andersen, T.G. and Lund, J. (1996b). Stochastic volatility and Mean Drift in the Short Term Interest Rate Diffusion: Sources of Steepness, Level and Curvature in the Yield Curve". *Manuscript.*

Anderson, T.W. (1971). *The Statistical Analysis of Time Series.* Wiley, New York.

Box, G. (1988). Signal-to-noise ratios, performance criteria, and transformations. *Technometrics*, **30**, 1-17.

Carroll, R. and Ruppert, D. (1988). *Transformations and Weighting in Regression.* Chapman & Hall, London.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of U.K. inflation. *Econometrica*, **50**, 987-1008.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.*, **21**, 196–216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**, 2008–2036.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc. B*, **57**, 371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625-633.

Gallant, A. R. and Tauchen, G. (1995). Estimation of continuous time models for stock returns and interests rates. (*A preprint.*)

Müller, H.G. and Stadtmüller U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610-625.

Müller, H.G. and Stadtmüller U. (1993). On variance function estimation with quadratic forms. *J. Statist. Plann. Inf.* **35**, 213-231.

Härdle, W. and Tsybakov, A. (1996). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, forthcoming.

Hall, P. and Carroll, R.J. (1989). Variance function estimation in regression: the effect of estimation of the mean. *J. Roy. Statist. Soc.* **B**, **51**, 3-14.

Hall, P., Kay, J. and Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521-528.

Hjellvik, V., Yao, Q. and Tjøstheim, D. (1995). Linearity testing using local polynomial approximation. (Submitted.)

Hastie, T.J. and Loader, C. (1993). Local regression: automatic kernel carpentry (with discussion). *Statist. Sci.*, **8**, 120–143.

Huang, L.S. (1995). On nonparametric estimation and goodness-of-fit. Ph.D. Dissertation, Department of Statistics, University of North Carolina at Chapel Hill.

Neumann, M.H. (1994). Fully data-driven nonparametric variance estimators. *Statistics*, **25**, 189-212.

Peligrad, M. (1986). Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables. *Dependence in Probability and Statistics*, Ed. E. Eberlein and M.S. Taqqu. Birkhäuser, Boston, 193-223.

Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.*, **12**, 1215-1230.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, **90**, 1257–1270.

Ruppert, D. and Wand, M.P. (1994). Multivariate weighted least squares regression. *Ann. Statist.*, **22**, 1346–1370.

Ruppert, D., Wand, M.P., Holst, U. and Hössjer, O. (1996). Local Polynomial Variance Function Estimation. *Manuscript.*

Schmidt, G., Mattern, R. and Schüler, F. (1981). Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under effects of impact. EEC Research Program on Biomechanics of Impacts. Final Report Phase III, Project 65, Institut für Rechtsmedizin, Universität Heidelberg, Germany.

Yao, Q. and Tong, H. (1994). Quantifying the influence of initial values on nonlinear prediction. *J. Roy. Statist. Soc.* **B**, **56**, 701-725.

Yao, Q. and Tong, H. (1996). Nonparametric estimation of ratios of noise to signal in stochastic regressions. (Submitted.)

Yoshihara, K. (1976). Limiting behaviour of U-statistics for stationary absolutely regular processes. *Z. Wahr. v. Gebiete*, **35**, 237-252.

Appendix: Proof of Theorem 1

We use the same notation as in §2. In the sequel, $\hat{m}(.)$ denotes the local linear estimator derived from (2.2). We always assume that conditions (C1) — (C5) hold. We call that $B_n(x) = B(x) + o_p(b_n)$ (or $O_p(b_n)$) uniformly for $x \in G$ if

$$\sup_{x\in G} |B_n(x) - B(x)| = o_p(b_n) \text{ (or } O_p(b_n)),$$

and $a_n \sim b_n$ if $a_n/b_n \to 1$. The proof is based on the following lemma which follows from Lemma 2 of Yao and Tong (1996) directly.

Lemma 1. Let $G \subset \{p(x) > 0\}$ be a compact subset. As $n \to \infty$, uniformly for $x \in G$,

$$\hat{\sigma}^2(x) - \sigma^2(x) = \left\{ \frac{1}{nh_1 p(x)} \sum_{i=1}^n W\left(\frac{X_i - x}{h_1}\right) \{\hat{r}_i - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\} \right\} \{1 + o_p(1)\}, \quad \text{(A.1)}$$

$$\begin{aligned} \hat{m}(x) - m(x) &= \left\{ \frac{1}{nh_2 p(x)} \sum_{i=1}^n K\left(\frac{X_i - x}{h_2}\right) \{Y_i - m(x) - \dot{m}(x)(X_i - x)\} \right\} \{1 + o_p(1)\} \quad \text{(A.2)} \\ &= \frac{1}{nh_2 p(x)} \sum_{i=1}^n \sigma(X_i)\epsilon_i K\left(\frac{X_i - x}{h_2}\right) + \frac{h_2^2 \sigma_K^2}{2} \ddot{m}(x) + o_p\left(\frac{1}{\sqrt{nh_2}} + h_2^2\right). \quad \text{(A.3)} \end{aligned}$$

Proof of Theorem 1. Note that

$$\begin{aligned} \hat{r}_i &= \{Y_i - \hat{m}(X_i)\}^2 = \{\sigma(X_i)\epsilon_i + m(X_i) - \hat{m}(X_i)\}^2 \\ &= \sigma^2(X_i)\epsilon_i^2 + 2\sigma(X_i)\epsilon_i\{m(X_i) - \hat{m}(X_i)\} + \{m(X_i) - \hat{m}(X_i)\}^2. \end{aligned}$$

It follows from (A.1) that

$$\hat{\sigma}^2(x) - \sigma^2(x) = \{I_1 + I_2 - I_3 + I_4\}\{1 + o_p(1)\},$$

where

$$\begin{aligned} I_1 &= \frac{1}{nh_1p(x)}\sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right)\{\sigma^2(X_i) - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\}, \\ I_2 &= \frac{1}{nh_1p(x)}\sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right)\sigma^2(X_i)(\epsilon_i^2 - 1), \\ I_3 &= \frac{2}{nh_1p(x)}\sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right)\sigma(X_i)\epsilon_i\{\hat{m}(X_i) - m(X_i)\}, \\ I_4 &= \frac{1}{nh_1p(x)}\sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right)\{\hat{m}(X_i) - m(X_i)\}^2. \end{aligned} \tag{A.4}$$

In the sequel, we will establish:

(a) $I_1 = \frac{1}{2}h_1^2\ddot{\sigma}^2(x)\sigma_W^2 + o_p(h_1^2)$,

(b) $\sqrt{nh_1}I_2 \xrightarrow{d} N(0,\ \sigma^4(x)\lambda^2(x)\int W^2(t)dt/p(x))$,

(c) $I_3 = o_p(h_1^2 + h_2^2)$,

(d) $I_4 = o_p(h_1^2 + h_2^2)$.

It is easy to see that the theorem follows from (a) — (d) directly.

It is easy to see that (a) follows from a Taylor's expansion, and a direct application of the ergodic theorem. Conditions (C2) and (C3) imply that $E\{W\left(\frac{X_i-x}{h_1}\right)\sigma^2(X_i)(\epsilon_i^2-1)\}^{2+\delta/2} < \infty$. Note that the condition of absolutely regular implies α-mixing with $\alpha(j) \le \beta(j)$. By (C4) and Theorem 1.7 of Peligrad (1986), I_2 is asymptotically normal with mean 0 and variance σ_*^2/nh_1 when $\delta > 0$, where

$$\begin{aligned} \sigma_*^2 &= \frac{1}{h_1}E\left\{W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\epsilon^2-1)\right\}^2 \\ &+ \frac{1}{h_1}\sum_{i=2}^{n} E\left\{W\left(\frac{X_1-x}{h_1}\right)\frac{\sigma^2(X_1)}{p(X_1)}(\epsilon_1^2-1)W\left(\frac{X_i-x}{h_1}\right)\frac{\sigma^2(X_i)}{p(X_i)}(\epsilon_i^2-1).\right\} \end{aligned} \tag{A.5}$$

It is easy to see that the first term in the above expression converges to $\sigma^4(x)\lambda^2(x)\int W^2(t)dt/p(x)$. Note that $E\left\{W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\epsilon^2-1)\right\} = 0$, $E\left|W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\epsilon^2-1)\right|^{1+\delta} = O(h_1)$, and for any $i \ge 2$,

$$E\left\{W\left(\frac{X_1-x}{h_1}\right)\frac{\sigma^2(X_1)}{p(X_1)}(\epsilon_1^2-1)W\left(\frac{X_i-x}{h_1}\right)\frac{\sigma^2(X_i)}{p(X_i)}(\epsilon_i^2-1)\right\}^{1+\delta} = O(h_1^2).$$

It follows from (C4) and Lemma 1 Yoshihara (1976) that the absolutely value of the second term in (A.5) is bounded by $ch_1^{(1-\delta)/(1+\delta)}\sum_{j=1}^{n-1}\beta^{\frac{\delta}{1+\delta}}(j) = o(1)$. Hence (b) holds when $\delta > 0$. In the case that $\delta = 0$, the asymptotic normality follows from a simple application of the standard small-block and large block arguments (see Remark 1).

Below we prove (c) for the case that $\delta > 0$. The case with $\delta = 0$ can be dealt in a more direct and simpler way. Note that $W(.)$ has bounded support, contained in the inverval $(-s_w, s_w)$, say. Therefore in the summation on the RHS of (A.4), only those terms with $X_i \in (x - h_2 s_w, x + h_2 s_w)$ might not be 0. It follows from (A.2) that

$$\begin{aligned} I_3 &\sim \frac{2}{n^2h_1h_2p(x)}\sum_{i,j=1}^{n} W\left(\frac{X_i-x}{h_1}\right)K\left(\frac{X_j-X_i}{h_2}\right)\frac{\sigma(X_i)\epsilon_i}{p(X_i)}\{\epsilon_j\sigma(X_j)+m(X_j) \\ &\qquad -m(X_i)-\dot{m}(X_i)(X_j-X_i)\} \\ &= \frac{2}{n^2h_1h_2p(x)}\sum_{1\le i<j\le n}\varphi_{ij}+O_p(\frac{1}{nh_2}), \end{aligned}$$

where $\varphi_{ij} = \psi_{ij} + \psi_{ji}$, and

$$\psi_{ij} = K\left(\frac{X_j-X_i}{h_2}\right)W\left(\frac{X_i-x}{h_1}\right)\frac{\sigma(X_i)\epsilon_i}{p(X_i)}\{\epsilon_j\sigma(X_j)+m(X_j)-m(X_i)-\dot{m}(X_i)(X_j-X_i)\}.$$

Performing Hoeffding's projection decomposition of U-statistic, we express

$$I_3 \sim \frac{2}{n^2h_1h_2p(x)}\sum_{1\le i<j\le n}\{\varphi_{ij}-\varphi_i-\varphi_j\}+\frac{2}{nh_1h_2p(x)}\sum_{i=1}^{n}\varphi_i+O_p(\frac{1}{nh_2}), \tag{A.6}$$

where

$$\begin{aligned} \varphi_i &= \frac{\sigma(X_i)\epsilon_i}{p(X_i)}W\left(\frac{X_i-x}{h_1}\right)\int K\left(\frac{z-X_i}{h_2}\right)\{m(z)-m(X_i)-\dot{m}(X_i)(z-X_i)\}p(z)dz \\ &= h_2^3\sigma_K^2\sigma(X_i)\epsilon_i W\left(\frac{X_i-x}{h_1}\right)\ddot{m}(X_i)+o_p(h_2^3). \end{aligned}$$

By (C2), (C4) and Theorem 1.7 of Peligrad (1986), the second term on the RHS of (A.6) is $O_p(h_2^2/\sqrt{nh_1})$. It follows from Lemma A(ii) of Hjellvik *et al.* (1996) that for any $\varepsilon_0 > 0$ and $\varepsilon > 0$,

$$\begin{aligned} & P\left\{\frac{1}{n(h_1h_2)^{(\frac{1}{1+\delta}-\varepsilon_0)/2}}\left|\sum_{i<j}(\varphi_{ij}-\varphi_i-\varphi_j)\right| > \varepsilon\right\} \\ \le\ & \frac{c(h_1h_2)^{\varepsilon_0}}{n^2}E\left\{\frac{1}{(h_1h_2)^{\frac{1}{2(1+\delta)}}}\sum_{i<j}(\varphi_{ij}-\varphi_i-\varphi_j)\right\}^2 = o\left((h_1h_2)^{\varepsilon_0}\right). \end{aligned}$$

Therefore, the first term on the RHS of (A.6) is $o_p\{n^{-1}(h_1h_2)^{-(\frac{1+2\delta}{1+\delta}+\varepsilon_0)/2}\}$. Thus

$$I_3 = o_p\left(\frac{1}{n(h_1h_2)^{(\frac{1+2\delta}{1+\delta}+\varepsilon_0)/2}}\right)+O_p\left(\frac{h_2^2}{\sqrt{nh_1}}\right)+O_p\left(\frac{1}{nh_2}\right).$$

Condition (C5) implies that all the three terms on the RHS of the above expression is of the order $o_p(h_1^2 + h_2^2)$ if we choose $\varepsilon_0 < (1+\delta)^{-1}$. This completes the proof of (c).

To prove (d), we apply asymptotic expression (A.3) directly. Using the similar argument as above, we can show that

$$I_4 = o_p\left(\frac{1}{n^{3/2}h_1h_2^2} + \frac{h_2}{nh_1} + \frac{1}{nh_2^2}\right) + O_p\left(\frac{h_2^3}{\sqrt{n}} + h_2^4\right) = o_p(h_1^2 + h_2^2).$$

Therefore, (d) holds.